Интернет-маркетинг и анализ данных поисковых запросов на основе алгоритмов машинного обучения в цифровой экономике

Плотников А.В., к.э.н, доцент кафедры менеджмента,
Пермский государственный аграрно-технологический университет имени академика Д.Н. Прянишникова, г. Пермь, Россия

Аннотация. В работе на основе применения алгоритмов машинного обучения построены прогнозные модели стоимости клика и сложности продвижения в контекстной рекламе Google Ads. В условиях несбалансированной выборки (количество значений одного класса значительно превосходит количество всех других классов), бинаризация таргетов позволила значительно повысить качество моделей. Переход к бинарной классификации также позволяет получить наглядное представление о качестве работы каждого классификатора. Наибольшее качество прогноза показала модель, основанная на деревьях решений. Для отображения качества прогнозов каждого из классификаторов использовали метод ROC-кривых.

Ключевые слова: цифровая экономика, контекстная реклама, поисковая оптимизация, поисковый маркетинг, интернет-маркетинг, машинное обучение, большие данные.

Internet marketing and data analysis of search queries based on machine learning algorithms in the digital economy

Plotnikov A.V., CSc., Associate Professor, Management Department, Perm State Agro-Technological University named after Academician D.N. Pryanishnikov, Perm, Russia

Annotation. In this paper, based on the use of machine learning algorithms, predictive models of cost per click and the complexity of promotion in contextual advertising Google Ads are built. In the conditions of unbalanced sampling (the number of values of one class significantly exceeds the number of all other classes), binarization of targets has significantly improved the quality of models. The transition to binary classification also provides a visual representation of the quality of each classifier. The model based on Random Forest method showed the highest quality of the forecast. To display the quality of forecasts of each of the classifiers, the ROC-curves method was used.

Keywords: digital economy, contextual advertising, search engine optimization, search engine marketing, Internet marketing, machine learning, big data.

Поисковые системы хранят данные о запросах пользователей – это уникальный набор данных, которые содержат информацию об интересах и намерениях пользователей и является показателем потребительского интереса к языку программирования HTML.

Многие исследования показывают, что частота поисковых запросов с конкретными словами могут быть предикторами социально-экономических явлений, связанных с потребительским поведением. Так анализ поисковых запросов пользователей в поисковых системах может применяться и при оценке риска дефолта по ипотеке [1]. Точный поисковый запрос помогает нам увидеть информацию, которую пользователь получил возможно после, ввода периферийных поисковых запросов. Как правило, короткие и неоднозначные поисковые запросы не всегда могут ответить релевантным ответом пользователю. Имея веб-адрес результатов, которые были просмотрены, мы можем извлечь дополнительную информация о точных потребностях пользователя [2].

Имеются собранные данные – поисковые запросы и их характеристики. Основная маска запроса: html. География: США. Период сбора данных август

2018г. Цель исследования предсказать данные ключевых слов по полям: «CPC», «Competition in PPC», difficulty алгоритмами машинного обучения XGBoost [3], Random forest [4], Lasso Regression [5].

Расшифровка данных (данные собраны на основе SaaS Serpstat) [6]:

The number of queries – количество запросов в поисковой системе Google;

CPC – cost per click – стоимость перехода в контекстной рекламе Google Ads;

Competition in PPC – показатель конкуренции в контекстной рекламе, где 0 – нет конкуренции, а 100 – слишком высокая конкуренция по пользовательскому запросу;

region_queries_count_phrase – количество региональных запросов (нет данных);

Keyword length – количество слов в запросе;

Difficulty – показатель конкуренции в органической выдаче в поисковой системе.

Таблица 1

Описательная статистика массива данных

	The number of queries	**CPC**	**Competition in PPC**	**Results found**	**region_queries_count_phrase**	**Keyword length**	**difficulty**
count	40000	40000	40000	4.000000e+04	40000	40000	5860
mean	119.886250	0.527904	5.812475	1.576471e+08	0.0	4.103225	24.915718
std	986.434139	2.547438	14.206698	8.303855e+08	0.0	1.598901	20.823711
min	20.000000	0.000000	0.000000	0.000000e+00	0.0	1.000000	0.000000
25%	30.000000	0.000000	0.000000	9.030000e+04	0.0	3.000000	7.070000
50%	50.000000	0.000000	0.000000	6.400000e+06	0.0	4.000000	20.150000
75%	90.000000	0.000000	4.000000	4.730000e+07	0.0	5.000000	37.465000

Из таблицы 1 можно увидеть, что, поле difficulty имеет много пропусков (count, т.е. количество значений для него 5760, против 40000 у других полей). Также можно получить представление о распределении данных в поле – есть минимум, максимум, среднее арифметическое, стандартное отклонение и распределение по квантилям.

Анализируем матрицу корреляции (рис. 1), обращаем внимание на относительно высокую связь между двумя целевыми столбцами: CPC и Competition in PPC.

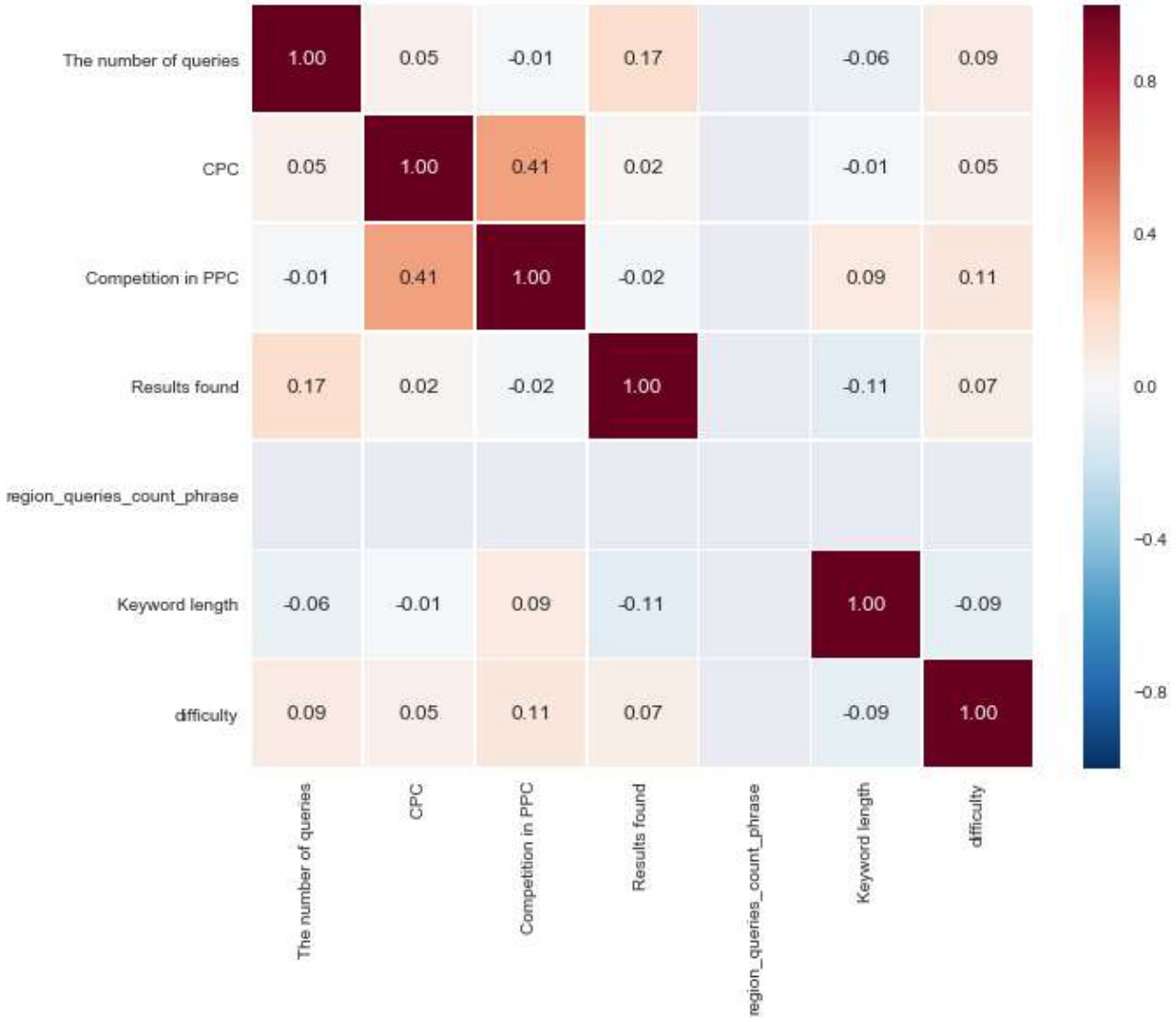

Рис. 1 – Корреляционная матрица

Затем рассмотрим каждую из labels отдельно:

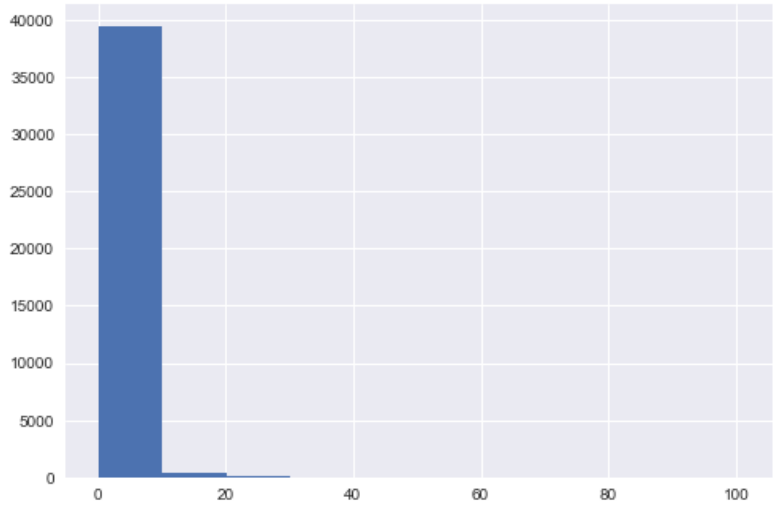

Рис. 2 – Распределение CPC (9,2% известных значений)

Рисунок 2 – это гистограмма, т.е. график, показывающий частоту попадания значений поля в интервалы. Из него видно, что основная часть значений лежит в интервале от 0 до 10. А 9% взялись из строчки ниже. Мы делим количество значений поля больших чем 0 на общее количество значений поля и получаем 0,092 или 9,2%. Если поле содержит только 9,2% (0,092) значений, отличных от нуля. В этой связи целесообразно бинаризировать поле, объединив значения в два класса «0» и «не 0».

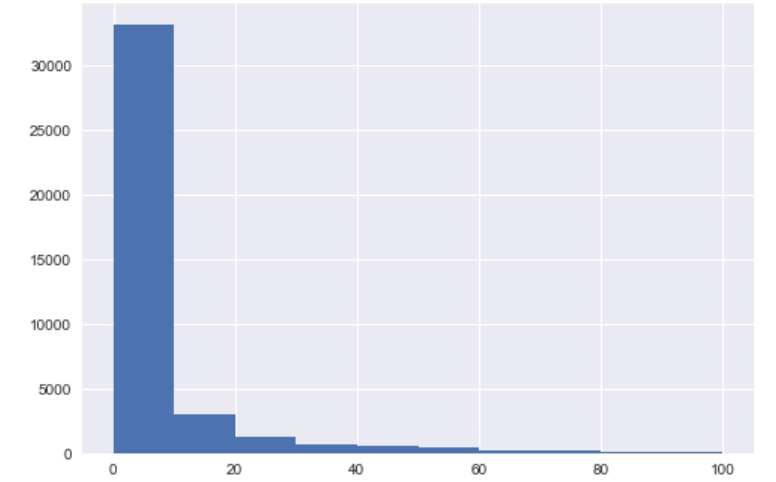

Рис. 3 – Распределение Competition in PPC (36,9% известных значений)

Поле содержит только 36,9 % значений, отличных от нуля. В этой связи целесообразно также бинаризировать поле, объединив значения в два класса «0» и «не 0».

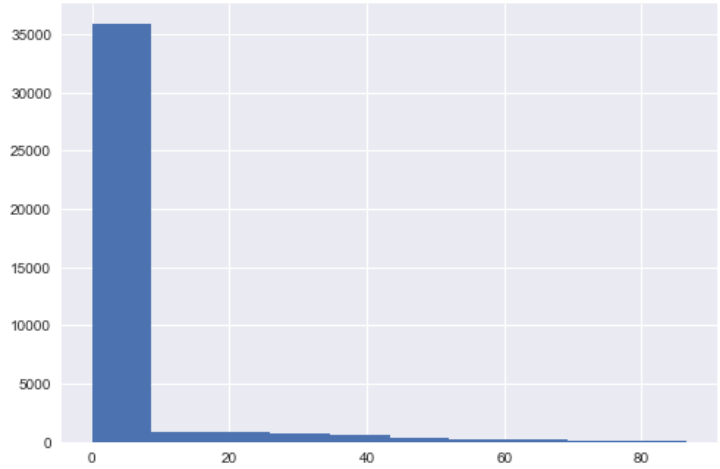

Рис. 4 – Распределение показателя Difficulty (13,8% известных значений)

Обращаем внимание, что поле «difficulty» содержит 5860 пустых (отсутствующих значений типа NaN). Заменяем их как планировали на минимальное значение, т.е. 0. Бинаризуем значения.

Строим тепловую матрицу корреляции и проверяем, что созданные labels близки к исходным (рис. 5)

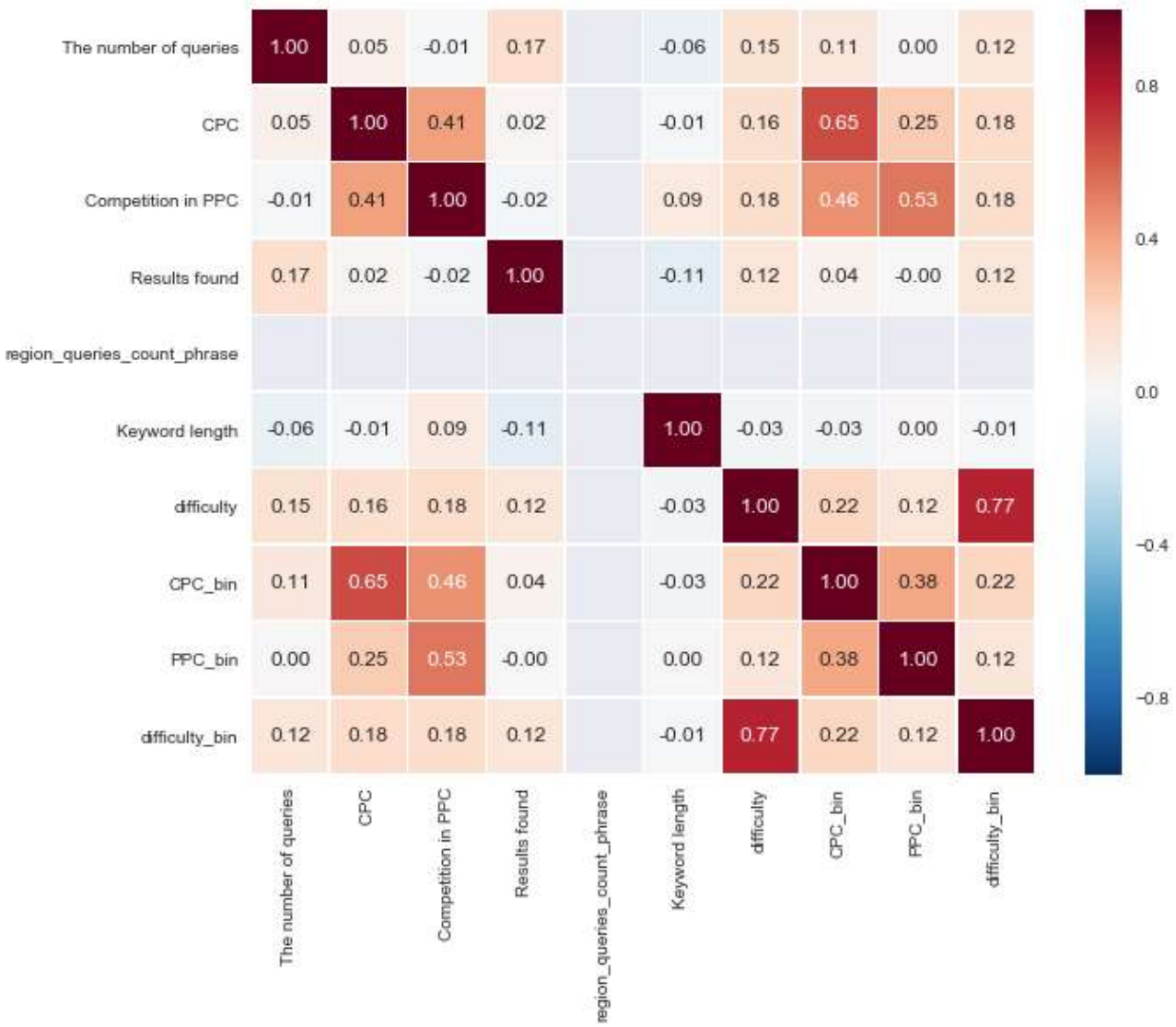

Рис. 5 – Корреляционная матрица, включая бинарные значения

Для использования машинного обучения слова или важно перевести в набор цифровых признаков (векторизовать), который его характеризует. Сделать это можно несколькими способами (по сути, очень близкими) один из которых tf-idf (TF — term frequency, IDF — inverse document frequency) [7]. Другой вариант – использовать готовые модели от rusvectores, например, там используется word2vec. В нашем случае, word2vec применять не практично, т.к. речь идет не о связанном тексте, а о, по сути, простом наборе слов в запросе.

Проводим векторизацию текстовых значений в цифровые методом Tf-idf, в результате получая: (40000, 5438) – это размерность массива после векторизации текста. 40 000 значений (по количеству запросов) как было, так и осталось, а вот вектор описывающий каждый запрос состоит из 5438 элементов. Этого количества будет достаточно для применения метода Random forest.

Задаем первое таргетное поле 'CPC_bin' и проводим обучение моделей.

Модель линейной регрессии с регуляризацией Лассо. Test Accuracy: 0.15;

Модель регрессора Random forest. Test Accuracy: 0.69;

Модель градиентного бустинга (XGBoost). Test Accuracy: 0.43.

Для отображения качества прогнозов каждого из классификаторов будем использовать метод ROC-кривых [8]. Для каждого классификатора просчитывается площадь под кривой. Единице соответствует классификатор со 100% точностью прогноза. Регрессор выдает непрерывные значения от 0 до 1. В зависимости от выбора порога отнесения предсказания к 0 или 1 (0,5 или 0,55 или 0,3, например) будет зависеть точность и полнота предсказаний модели. Порог выбирается в зависимости от задачи (что лучше: пропустить истинно-положительный результат или предсказать ложноположительный). Все это как раз очень наглядно демонстрирует ROC-кривая.

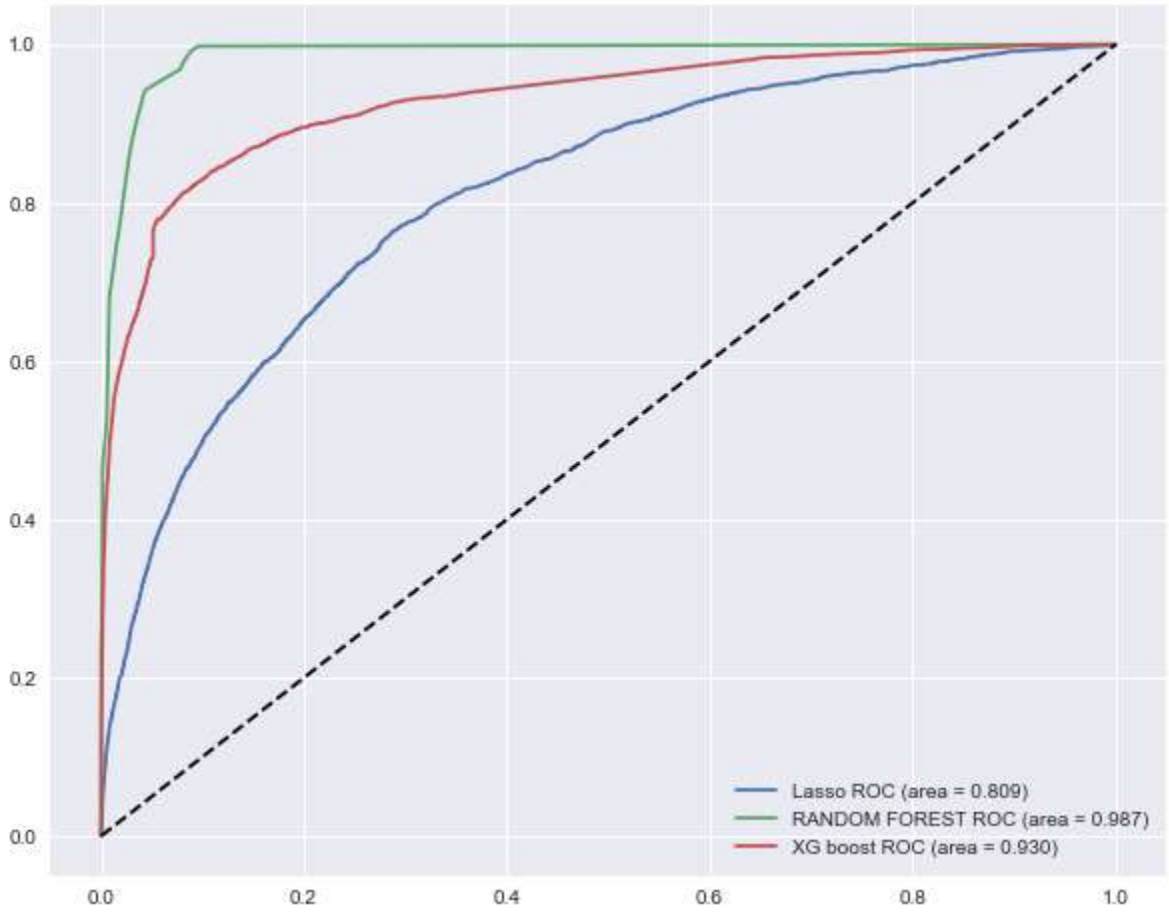

Рис. 6 – Оценка качества прогноза по таргетному полю 'CPC_bin' методом ROC-кривых

Как можно видеть из графика наилучшее качество прогноза дает Random forest классификатор.

Задаем следующее таргетное поле «PPC_bin» и тренируем те же модели

Test Accuracy (регрессия с регуляцией Лассо): 0.15

Test Accuracy (Random forest): 0.69

Test Accuracy (Модель градиентного бустинга): 0.43

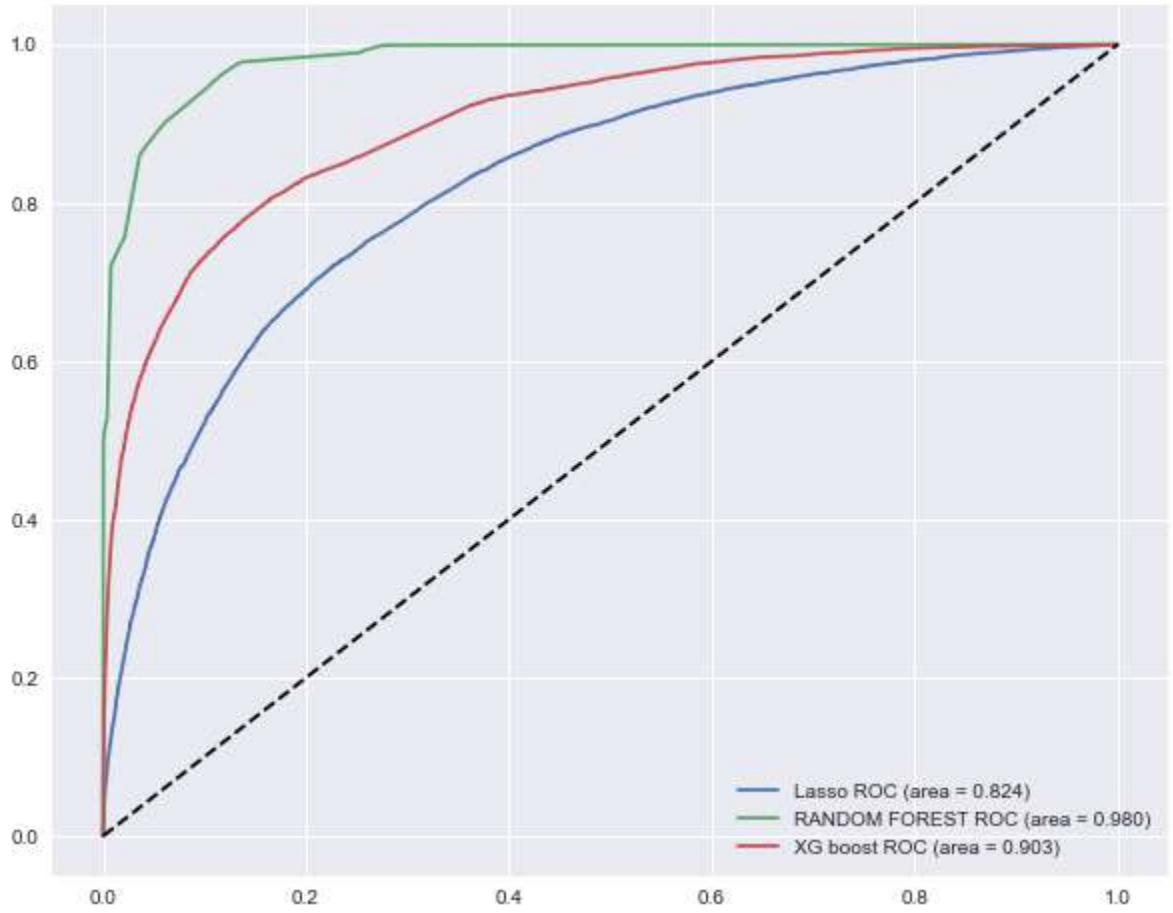

Рис. 7 – Оценка качества прогноза по таргетному полю 'PPC_bin' методом ROC-кривых

Задаем поле «difficulty_bin» в качестве таргета и тренируем модели

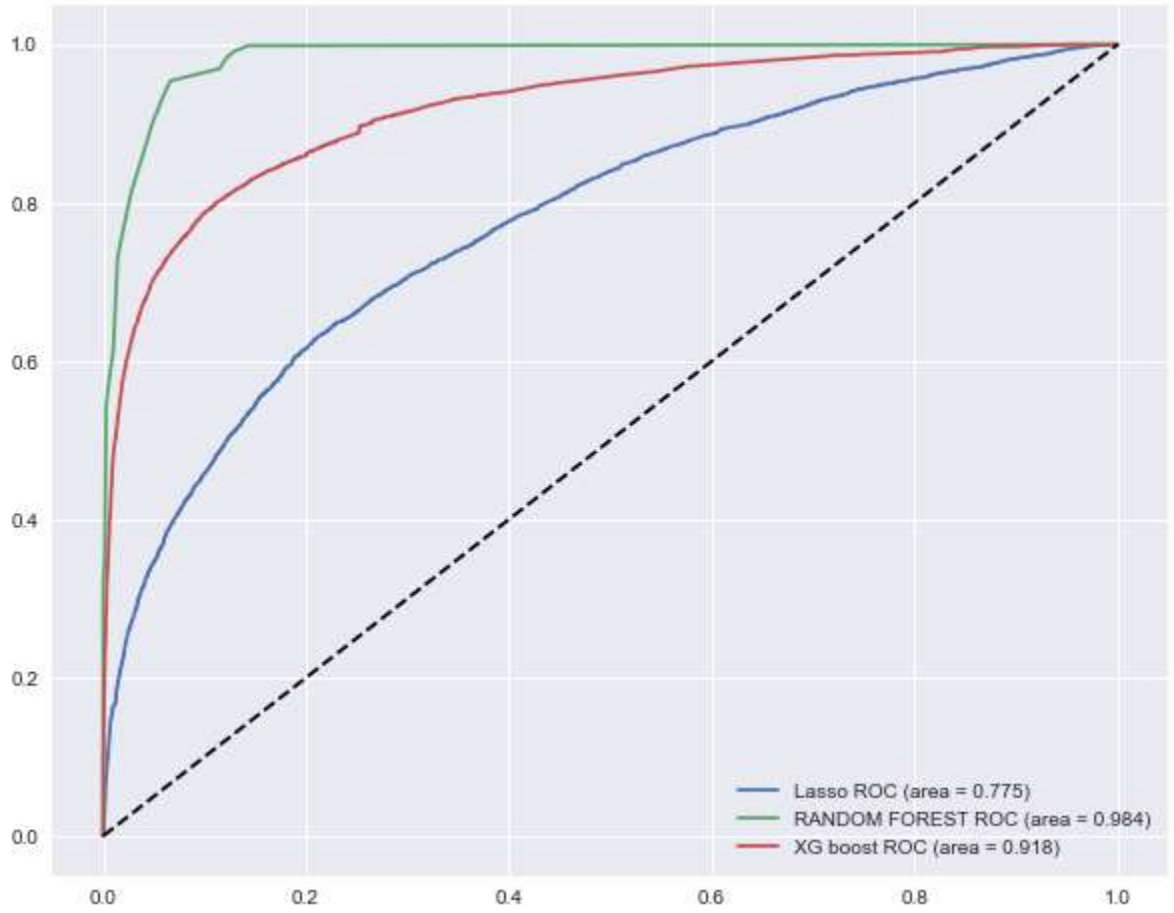

Рис. 8 – Оценка качества прогноза по таргетному полю 'difficulty_bin' методом ROC-кривых

В условиях несбалансированной выборки (количество значений одного класса значительно превосходит количество всех других классов), бинаризация таргетов позволила значительно повысить качество моделей. Переход к бинарной классификации также позволяет получить наглядное представление о качестве работы каждого классификатора. Наибольшее качество прогноза показала модель, основанная на деревьях решений. Возможно, линейная модель при дальнейшем уменьшении коэффициента регуляризации, смогла бы приблизиться к ней по точности прогноза, однако потребовала бы значительного времени на расчет.

Исследование выполнено при финансовой поддержке РФФИ и Пермского края в рамках научного проекта № 18-410-590007

Библиографический список

1. Lin Li, Luo Zhong, Guandong Xu, Masaru Kitsuregawa A feature-free search query classification approach using semantic distance // Expert Systems with Applications, Volume 39, Issue 12, 2012, pp. 10739-10748, https://doi.org/10.1016/j.eswa.2012.02.191.

2. Mondria J., Wu T., Zhang Y. The determinants of international investment and attention allocation: Using internet search query data //Journal of International Economics. – 2010. – Т. 82. – №. 1. – С. 85-95.

3. Chen T., Guestrin C. Xgboost: A scalable tree boosting system //Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – ACM, 2016. – С. 785-794.

4. Liaw A. et al. Classification and regression by randomForest //R news. – 2002. – Т. 2. – №. 3. – С. 18-22.

5. Meier L., Van De Geer S., Bühlmann P. The group lasso for logistic regression //Journal of the Royal Statistical Society: Series B (Statistical Methodology). – 2008. – Т. 70. – №. 1. – С. 53-71.

6. SaaS Serpstat [Web] serpstat.com

7. Shouzhong T., Minlie H. Mining microblog user interests based on TextRank with TF-IDF factor // The Journal of China Universities of Posts and Telecommunications. – 2016. – Т. 23. – №. 5. – С. 40-46.

8. Beck J.R., Shultz E.K. The use of relative operating characteristic (ROC) curves in test performance evaluation // Archives of pathology & laboratory medicine. – 1986. – Т. 110. – №. 1. – С. 13-20.